

PRELIMINARY VERSION FOR HANDOUT AT THE
1993 JOINT STATISTICAL MEETINGS

A LOOK AT A DAY OF DATA FROM THE TOPEX/POSEIDON GPS RECEIVER

J. Rodney Jee
Jet Propulsion Laboratory
California Institute of Technology

Introduction

Launched on 10 August, 1992, the Topex/Poseidon satellite is a joint project of NASA and the French space agency CNES. The purpose of this Earth orbiting altimetric satellite (for brevity, referred to as Topex) is to enable monitoring of the ocean topography to higher levels of accuracy than previously achieved. Essential to obtaining this greater level of accuracy is the use of improved navigational systems which can yield position estimates for Topex good to a few centimeters. One of the navigational systems utilized by Topex is the Global Positioning System (GPS)—a network of navigational beacons that has become widely used in many areas beside spacecraft navigation. Topex is one of the first satellites to utilize a high precision dual frequency GPS receiver and the status of the use of the GPS for Topex is that of an experiment rather than as the primary operational system for navigation. In addition to the on-board receiver, the navigation experiment requires a network of ground based GPS receivers to collect data simultaneously with Topex.

The focus of this talk is on the data pre-processing (specifically, the data editing) that must be performed on the GPS data from Topex before it is passed to the orbit determination part of the software. The editing is done by a program that is part of the GPS Data Processing Facility (GDPF)—a set of software designed at JPL for this navigation experiment. It should come as no surprise that some form of data editing must be performed initially. The main goal of the data editor is to relieve the orbit analyst of further need for data editing—especially in the orbit determination stage where problems arising from faulty data are more difficult and computationally expensive to diagnose. Numerous computer programs for editing GPS data are in use within the GPS user community and the data editor of the GDPF has incorporated some of the ideas of these programs while adding in some additional and improved procedures. As an integral part of its processing the GDPF data editor maintains a statistical summary of all processing that is performed. In addition it automatically generates a fairly complete set of diagnostic graphics in postscript form which may be printed or viewed on a computer monitor. A sample of this graphical output will be presented here along with some highlights of the workings of the data editor. A goal of this talk is to show the role that good graphics and statistics has in facilitating the processing and understanding of large amounts of data.

The Data: Names and Volume

The Topex Design Receiver (DR) has six channels to allow it to simultaneously track 6 GPS satellites (see Figure 1). These 6 satellites are optimally selected at regular intervals by the receiver from a constellation of over 24 (not all of which are in view at any instant).

Each of the 6 channels records a multivariate time series which I now describe briefly. For each GPS the pseudorange data type is available on carriers at 2 frequencies (1575.42GHz and 1227.6GHz, respectively, or about 0.19cm and 0.24cm) at the rate of 1 point per 10 seconds each. This data type is most naturally thought of as a measurement giving the time it takes the signal to travel from the GPS transmitter to the receiver. Its units are time, but it is common to convert pseudorange to units of length (by multiplying by the speed of light) which is done in the GDFP. These data have decimeter level precision and will be referred to as P1 and P2. For each GPS the phase data type is available at the same two frequencies as pseudorange but at the higher rate of once per second. This data type can be thought of as accumulated counts of cycles of the received signal. Multiplication by the appropriate wavelength converts these to units of length. These data have millimeter level precision and will be designated L1 and L2. The set of P1, P2, L1, and L2 data recorded by uninterrupted, continuous tracking of a particular GPS is referred to as a "pass" of data. Typically, the DR collects over 300 passes of data a day. Since the DR continuously collects phase data every second simple arithmetic shows that over 1 million points per day must be processed. The differential span of both the pseudorange and phase data during a pass can be about 8000 km. The need to retain better than millimeter precision in the data thus requires high precision arithmetic.

The Main Statistical Tools Used in Data Editing

All of the data are fit (per time series, per pass) with robust regression splines. The purpose in performing the spline fits are to provide a means to spot faulty data—outliers or other specific problems known to occur for the data. There are over 20 user specifiable parameters for controlling how the spline fits are done. At this point I list some of the parameters and their default values if applicable:

- (1) Constant knot spacing (about 10 data points per knot),
- (2) Constant knot repetition to control degree of derivative continuity (no repetition for maximum smoothness)
- (3) Constant polynomial degree
 - Cubic Spline for 1/sec phase data
 - Quartic Spline for 1/10 sec pseudorange data
- (4) Huber weighting function
- (5) Physics based pre-screening of obvious outliers to produce better initial least squares spline estimates (controlled by specifiable parameters).
- (6) IRLS algorithm for optimization (with specifiable convergence conditions)

Robust 2-D discriminant analysis is performed with

Median Absolute Deviation (MAD) based variances, and
 MAD based covariances via robust version of

$$\text{COV}(X,Y) = (\text{VAR}(X+Y) - \text{VAR}(X-Y))/4.$$

The S-PLUS Software from Statistical Sciences, Inc., provides

a Fortran interface for running subroutines of the data editor (which is in Fortran) with complete diagnostics and convenient access to intermediate results, graphical capabilities (X-windows and postscript), and

a flexible programming environment used to perform special analyses and prototyping of algorithms for Fortran implementation.

The Pseudorange Data and Problems

Figure 2 shows plots of a few hours of the pseudorange data from channel 5 of the DR. The raw (unedited) P1 are shown in the subplot on the left. The “ionospheric combination” P1 -P2 for the same time span is shown in the subplot on the right. The main problems with pseudorange are occasional outliers which are readily detected and deleted by checking the residuals from the spline fits made to each of P1 and P2 for each pass of data. The data editor has user specifiable parameters for controlling what constitutes large residuals (in terms of the MAD of the residuals in a pass) and deletes the data with large residuals from further processing. The data editor automatically generates more condensed versions of these plots for the entire amount of pseudorange data every day. The data analyst usually does not inspect these pseudorange plots unless some kind of unusual problem occurs such as happened on the day of this plot. On a few occasions the DR has had difficulty tracking the pseudorange with results as seen near 17.5 hours of the day in the subplot on the left. For this pass of data, the robust spline fitting algorithm reported non-convergence which is an unusual problem for the data editor. The data editor is programmed to delete the entire pass of data if a spline fit fails to converge which in the case of this pass is the appropriate action.

The Phase Data

Figure 3 shows plots of one pass of phase data in the linear combinations that are usually formed by GPS data analysts. The top plot shows the linear combination of phase data often designated LC ($\cong L1 + 1.54(L1 - 1.2)$). This is the only combination of the phase data which is used by the orbit determination part of the GDPF as the dual frequencies are for ionospheric calibration purposes (to be made more clear soon). Since the differential span of the LC is typically thousands of kilometers while phase problems are often sub-meter in level, plots of LC seldom reveal any problems discernible by the unaided eye. The bottom plot of Figure 3 shows the ionospheric combination 1.1-1.2. This linear combination is dominated by ionospheric effects and is of a scale as to allow one to sometimes see directly the level of the noise in the data. There are some noticeable “discontinuities” in the 1.1-1.2 plot of Figure 3. They are manifestations of what are known as cycle slips—the main problem the GPS data editors must contend with. It has turned out that the Topex-GPS Receiver rarely has cycle slips, so the data for Figure 3 is not chosen as a representative sample but rather as one of the few that does have the cycle slip problem.

The Cycle Slip Problem for Phase Data

As illustrated in Figure 3 the cycle slip problem results in discontinuities in the phase data. The sizes of the discontinuities are multiples of a half-integer when the data are in units of cycles in the original L1 and 1.2 (not in the LC and 1.1- 1.2 combinations).

The data editing software tackles the cycle slip problem by forming the divided differences of the data in time thereby converting the discontinuities into outliers. Then robust regression splines are fit to the rate(LC) and rate(L1-1.2) to produce residuals by which cycle slips are identified. See Figure 4 for time series plots of the rates of the phase. “The bottom plot of Figure 4 overlays the spline fit on the data values.

It is highly desirable to fix the cycle slips if possible. The data editing software fixes cycle slips by using two-dimensional discriminant analysis to resolve how many half-integers were jumped. Figure 5 shows a plot of the residuals from spline fits to the data of Figure 4 in a scatter-plot along with contours of equal probability centered at half-integer locations. Note that the residuals in $\text{rate}(\text{LC})$ and $\text{rate}(\text{L}1 - \text{L}2)$ have been transformed to units of cycles in $\text{L}1$ and $\text{L}2$. As usual in discriminant analysis the contours are based upon Gaussian assumptions but using robust covariance estimation as indicated earlier. The cycle slip identification algorithm proceeds by first determining the cycle slip center to which a point is closest. Then it checks that a point is within some user specified acceptance region to bound the probability of misclassification. If a point cannot be classified with very low probability of error (as specified by the user), then the cycle slip is simply marked as the beginning of a new pass. In this case all the cycle slips were properly identified and the fixed $\text{L}1$ - $\text{L}2$ combination is shown in Figure 6.

Statistics from the Data Editor

The data editor calculates and saves over 30 statistics for every pass of data it successfully processes. These statistics provide a means of monitoring the editor's performance and of monitoring the receiver's activities; unusual statistics could signal that processing is invalid or that some unanticipated problem slipped through the editor. An examination of these statistics for each of the 300 plus passes per day is the main way the editor's performance is validated before the edited data are passed on to additional processing. Of course, graphical representation of these statistics greatly aids in understanding. For each day of data processing the editor automatically produces over 10 postscript files of graphics (most with multiple plots) of these statistics. Figure 1 is a simplified version of one of these daily graphical summaries.

In the discussion on the pseudorange data, it was mentioned that the pass which was not tracked properly by the DR caused the IRI .S algorithm to fail to converge. A code identifying the type of failure is a statistic kept by the editor in the summary file, and it allowed me to match the graphic with the pass producing it.

As another example, Figure 7 shows a scatter-plot of the MADs of all the passes of residuals from the spline fits to LC and $\text{L}1$ - $\text{L}2$ phase combinations for 31 May 1993. A point with extreme statistics appears in this plot and signals that the pass should be inspected for some type of failure in the editing. The $\text{L}1$ - $\text{L}2$ combination for that pass is plotted in Figure 8. For a GPS data analyst this plot reveals some obvious problems with the data that had not been anticipated when designing the editor. The appropriate action for this pass is to delete all of the data and determine whether it is the receiver or the transmitter which is at fault. As it turned out, GPS 31 was experiencing hardware problems on that day.

The statistics in Figure 7 also give information about the level of the data noise in the phase data. If the outlying points in the scatterplot are ignored, then it is seen that the MADs of the difference LC spline residuals run between 3 and 10 mm. However, there is an apparent gap near the central region of the scatter. A plot of the same MADs versus the medians of the pseudoranges per pass shown in Figure 9 provides insight into the nature of the gap. The lower portion of the scatter shows a dependence of the noise level on the median of the pseudorange. This finding is expected since for the special geometry of Topex and the GPS constellation, the pseudorange is roughly a monotone function of the elevation angle and a dependence of noise levels on elevation angle has been observed in ground receivers. What is unexpected is that the upper band of points seem to be constant

in the MAD. This scatterplot prompted me to perform more detailed analyses which resulted in a finding that the bifurcation in noise levels is due to characteristics of the Topex DR itself rather than some other external or geometric factor.

SUMMARY

The main innovations in this data editor are the systematic use of robust statistical procedures, the formulation of the cycle slip problem as a statistical discrimination problem, and the systematic use of summary statistics and graphics to facilitate a quick assessment of editing results and receiver performance. Time limitations have permitted only brief descriptions of these parts of the data editor and some important details have not been discussed. Despite the omissions it is hoped that those unfamiliar with the GPS can gain some appreciation of the data pre-processing task and see how some of the developments in robust statistics should find a place in similar areas of automated processing.

ACKNOWLEDGEMENTS

The research in this presentation was earned out by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration. The parts of the GPS Data Processing Facility software described here have been developed by Vic Legerton, Bobby Williams, Joe Guinn, Chesley McColl, Tim Munson, and the author.

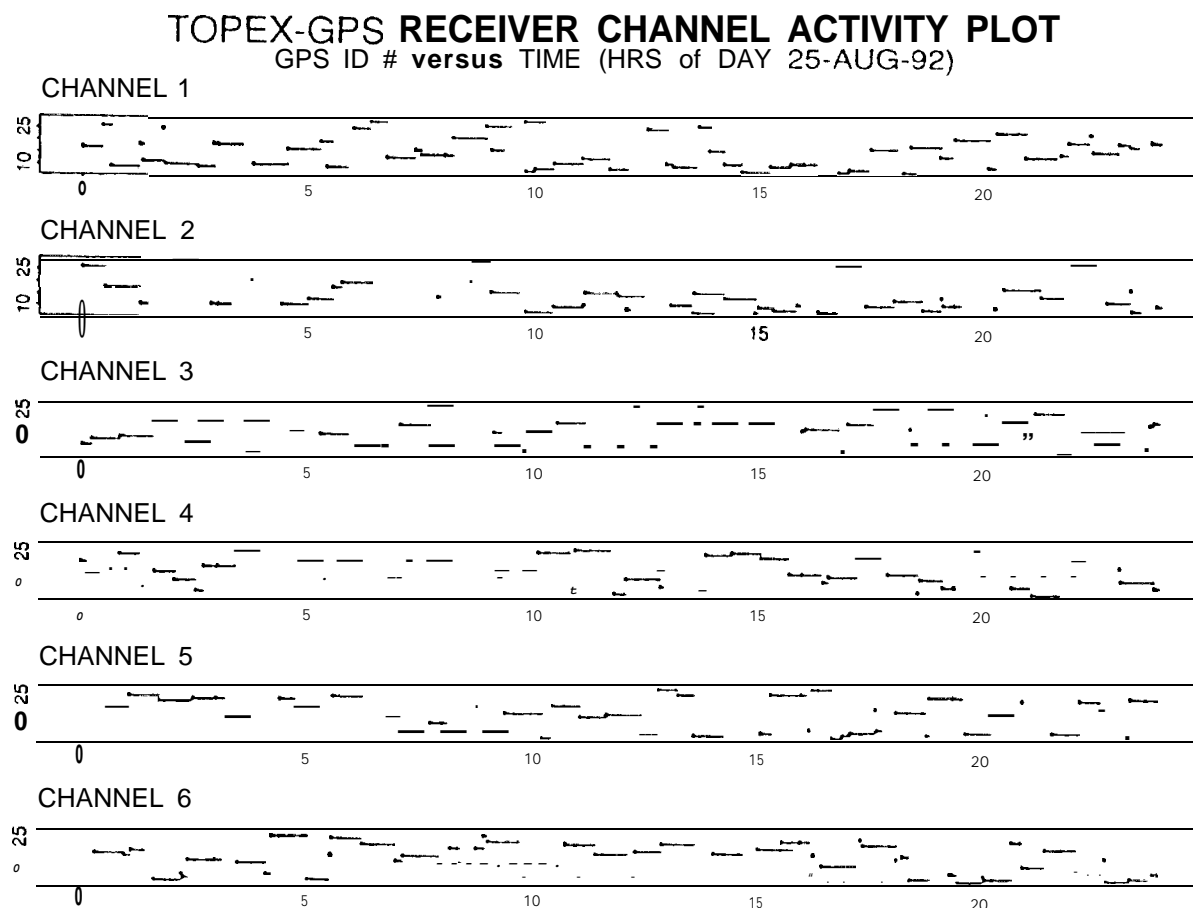


FIGURE 1: FIRST FULL DAY OF DATA FROM THE TOPEX-GPS RECEIVER

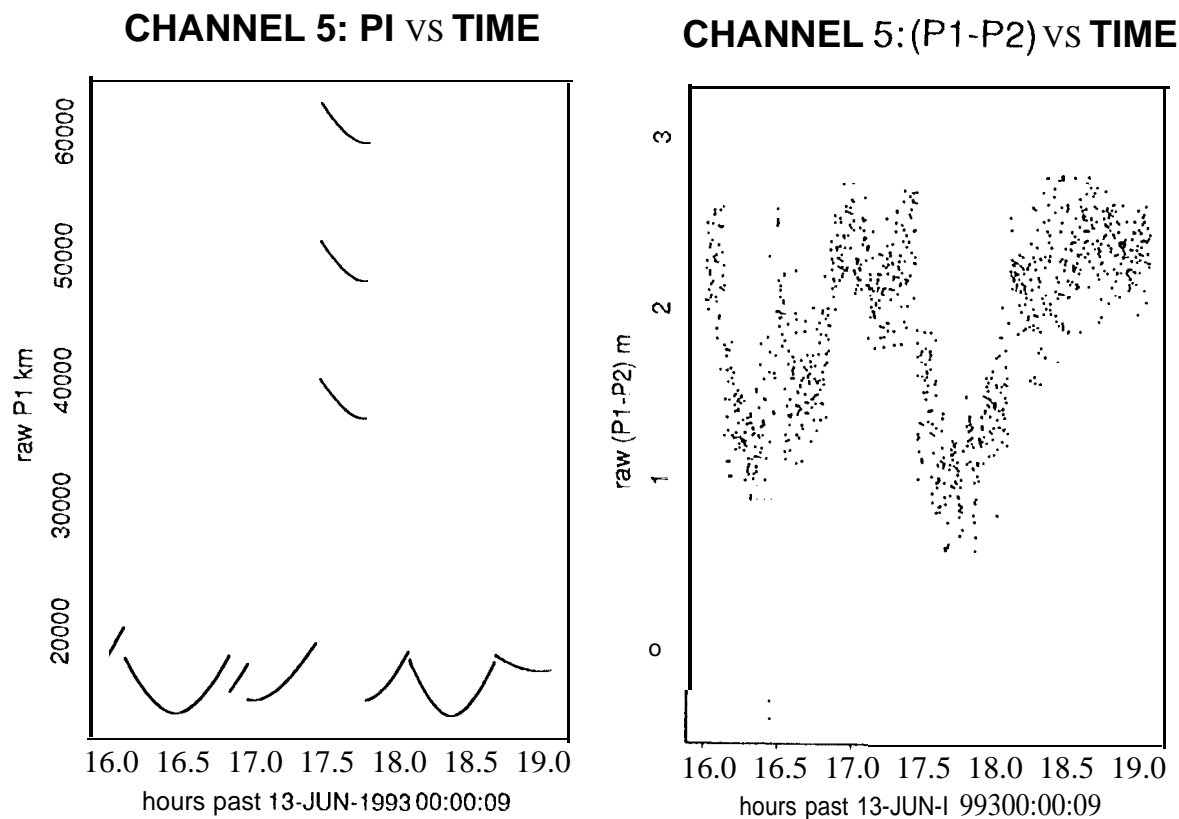


FIGURE 2: PSEUDORANGE FROM THE TOPEX GPS RECEIVER

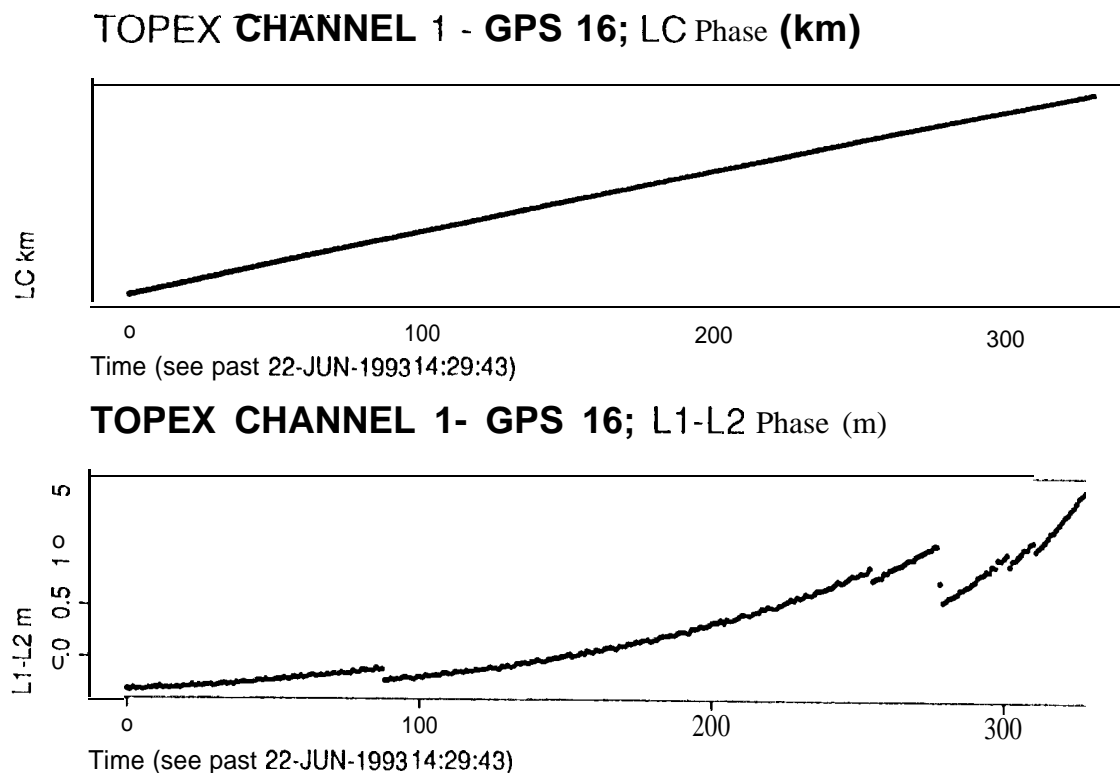


FIGURE 3: A PASS OF PHASE DATA with CYCLE SLIPS

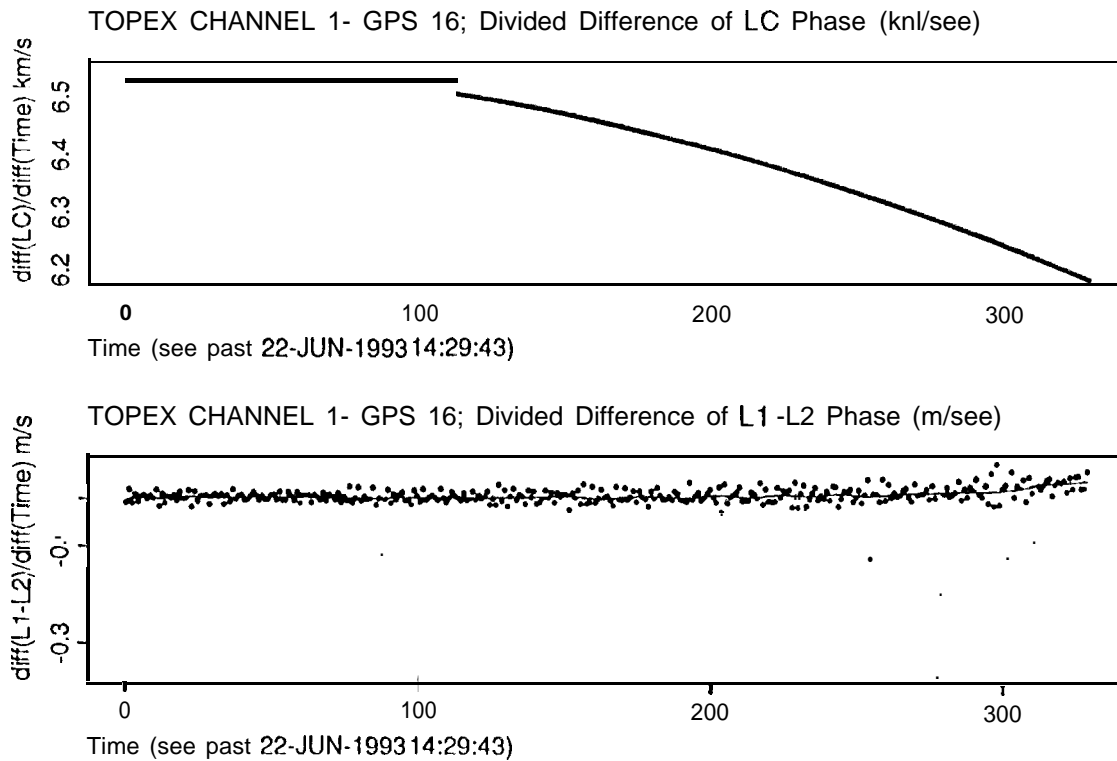


FIGURE 4: RATE of PHASE DATA and SPLINE FIT

TRANSFORMED RESIDUALS FROM SPLINE FITS
WITH CONTOURS BASED ON ROBUST COVARIANCE

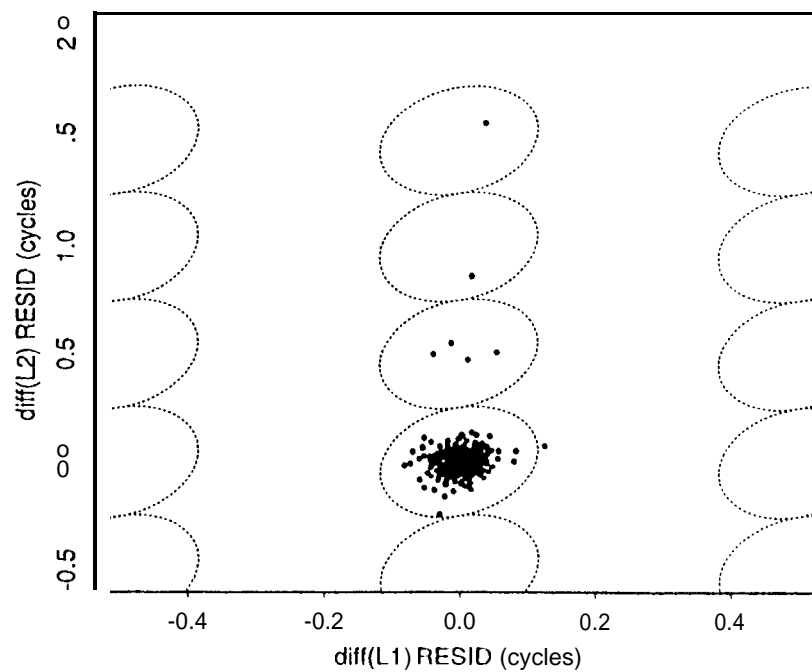


FIGURE 5: CYCLE SLIP DISCRIMINATION

TOPEX CHANNEL 1- GPS 16; L1-L2 FIXED PHASE

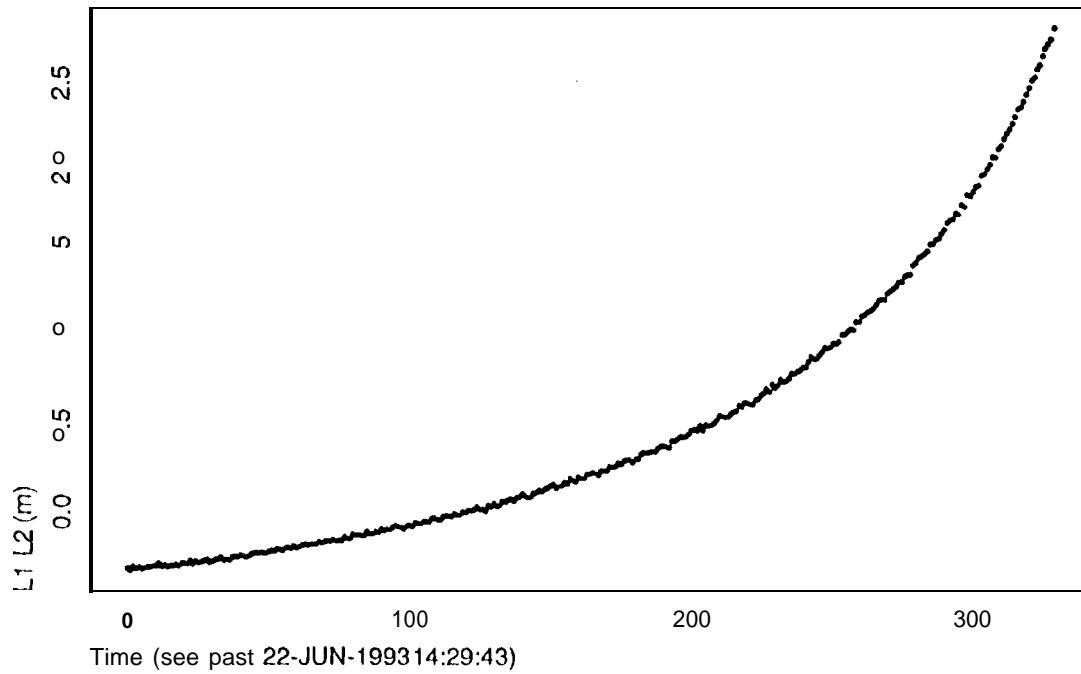


FIGURE 6: PHASE DATA CORRECTED FOR CYCLE SLIPS

TOPEX 31-MAY-93; STATISTICS OF SPLINE FITS TO LC AND L1-L2

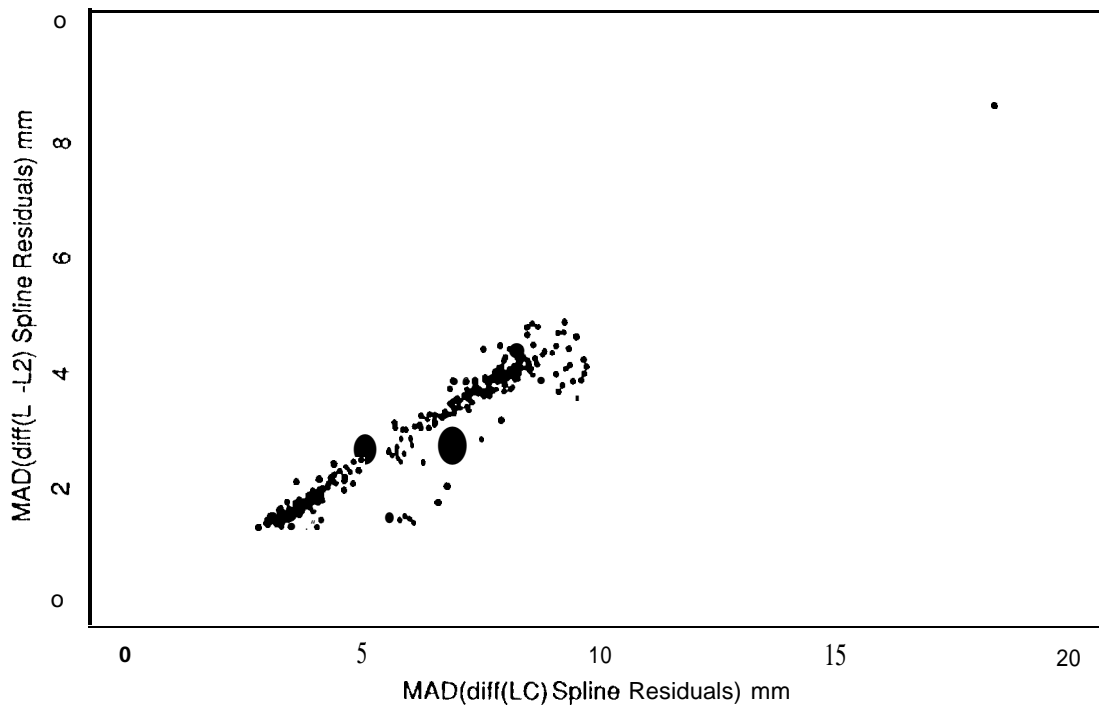


FIGURE 7: A DIFFERENT VIEW OF A DAY OF DATA

TOPEX CHANNEL 6 - GPS 31; LI-L2 PHASE DATA

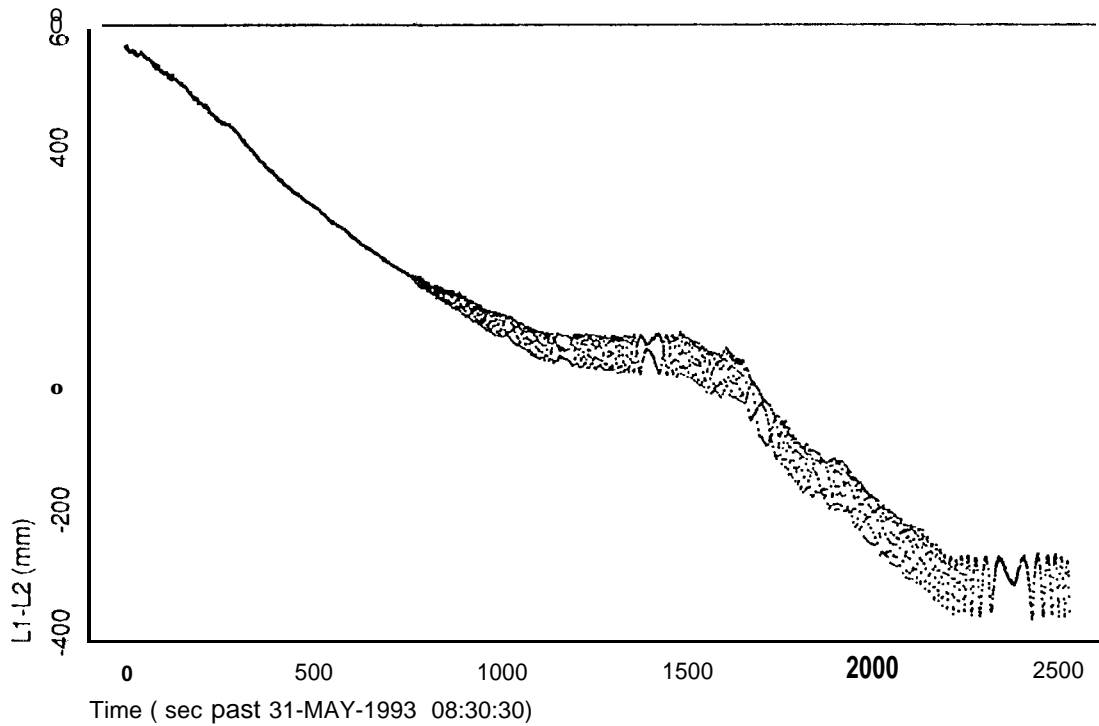


FIGURE 8: PASS WITH UNUSUAL STATISTICS

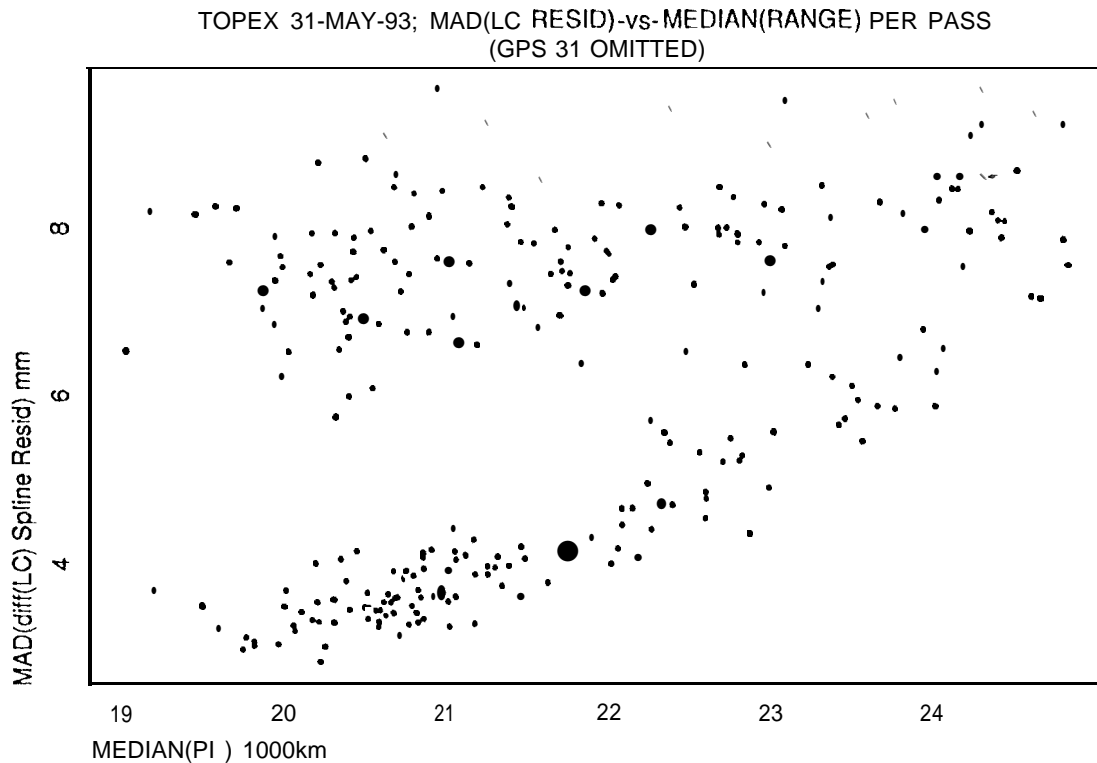


FIGURE 9: A BIFURCATION IN NOISE LEVELS